Generic comparison of protein inference engines

Claassen, M ; Reiter, L ; Hengartner, M O ; Buhmann, J M ; Aebersold, R

Abstract: Protein identifications, instead of peptide-spectrum matches, constitute the biologically relevant result of shotgun proteomics studies. How to appropriately infer and report protein identifications has triggered a still ongoing debate. This debate has so far suffered from the lack of appropriate performance measures that allow to objectively assess protein inference approaches.This study describes an intuitive, generic and yet formal performance measure and demonstrates how it enables experimentalists to select an optimal protein inference strategy for a given collection of fragment ion spectra. We applied the performance measure to systematically explore the benefit of excluding possibly unreliable protein identifications, such as single hit wonders. Therefore, we defined a family of protein inference engines, by extending a simple inference engine by thousands of pruning variants, each excluding a different specified set of possibly unreliable identifications. We benchmarked these protein inference engines on several datasets representing different proteomes and mass spectrometrical platforms. Optimally performing inference engines retained all high confidence spectral evidence, without posterior exclusion of any type of protein identifications.Despite the diversity of studied datasets consistently supporting this rule, other datasets might behave differently. In order to ensure maximal reliable proteome coverage for datasets arising in other studies, we advocate to abstain from rigid protein inference rules, like exclusion of single hit wonders, and instead to consider several protein inference approaches and to assess these with respect to the presented performance measure in the specific application context.

DOI: https://doi.org/10.1074/mcp.O110.007088

MCP Papers in Press. Published on November 4, 2011 as Manuscript O110.007088

Generic Comparison of Protein Inference Engines

Manfred Claassen[1,2,*], Lukas Reiter[2,3,4,*], Michael O. Hengartner[3], Joachim M. Buhmann[1], Ruedi Aebersold[2,5]

1) Department of Computer Science, ETH Zurich

2) Department of Biology, Institute of Molecular Systems Biology, ETH Zurich

3) Institute of Molecular Biology, University of Zurich

4) Biognosys AG, Zurich, Switzerland

5) Faculty of Science, University of Zurich

* contributed equally

Corresponding Author:

Prof. Ruedi Aebersold

Institute of Molecular Systems Biology

Wolfgang-Pauli-Str. 16, HPT E 78

ETH Zurich

CH-8093 Zurich

Phone: +41 44 633 31 70

Fax: +41 44 633 10 51

aebersold@imsb.biol.ethz.ch

Manfred Claassen

Department of Computer Science

Universitaetstrasse 6, CAB E 16

ETH Zurich

CH-8092Zurich

manfred.claassen@inf.ethz.ch

Lukas Reiter

Institute of Molecular Systems Biology, Department of Biology, Swiss Federal Institute of Technology (ETH) Zurich, Zurich, Switzerland.

Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland.

Current address: Biognosys AG, Zurich, Switzerland Wagistrasse 25, 8952 Schlieren reiter@biognosys.ch

Biognosys AG, Zurich, Switzerland

Wolfgang-Pauli-Str. 16, HPT C119

CH-8093 Zurich

lukas.reiter@biognosys.ch

Prof. Michael O. Hengartner

Institute of Molecular Biology

Winterthurerstrasse 190

University of Zurich

CH-8057 Zurich

michael.hengartner@molbio.uzh.ch

Prof. Joachim M. Buhmann

Department of Computer Science

Universittstrasse 6, CAB G 69.2

ETH Zurich

CH-8092Zurich

jbuhmann@inf.ethz.ch

Protein identifications, instead of peptide-spectrum matches, constitute the biologically relevant result of shotgun proteomics studies. How to appropriately infer and report protein identifications has triggered a still ongoing debate. This debate has so far suffered from the lack of appropriate performance measures that allow to objectively assess protein inference approaches.

This study describes an intuitive, generic and yet formal performance measure and demonstrates how it enables experimentalists to select an optimal protein inference strategy for a given collection of fragment ion spectra. We applied the performance measure to systematically explore the benefit of excluding possibly unreliable protein identifications, such as single hit wonders. Therefore, we defined a family of protein inference engines, by extending a simple inference engine by thousands of pruning variants, each excluding a different specified set of possibly unreliable identifications. We benchmarked these protein inference engines on several datasets representing different proteomes and mass spectrometrical platforms. Optimally performing inference engines retained all high confidence spectral evidence, without posterior exclusion of any type of protein identifications.

Despite the diversity of studied datasets consistently supporting this rule, other datasets might behave differently. In order to ensure maximal reliable proteome coverage for datasets arising in other studies, we advocate to abstain from rigid protein inference rules, like exclusion of single hit wonders, and instead to consider several protein inference approaches and to assess these with respect to the presented performance measure in the specific application context.

1 Introduction

A fundamental goal of mass spectrometry based proteomics is to determine the true protein composition of biological samples. Protein inference denotes the task of recovering the protein identities from the fragment ion spectra acquired in the course of shotgun proteomics experiments. Assessment of protein inference methods so far suffered from the lack of a generally applicable performance criterion that takes protein identification reliability into account. We extend the statistical validation framework Mayu [29] to define such a criterion and apply it to benchmark a family of prototypical protein inference approaches.

Protein inference is a task that arises in the context of shotgun proteomics experiments [2, 11]. In their simplest implementation, proteins are first extracted from their biological source, subjected to enzymatic digestion, yielding a complex peptide mixture that is analyzed by liquid chromatography tandem mass spectrometry (LC-MS/MS). Fragment ion spectra are acquired after stochastic or directed precursor ion selection [32]. More elaborate strategies augment this workflow by additional fractionation steps at the level of proteins/peptides before LC-MS/MS analysis. These steps give rise to a set of peptide fragment ion spectra that constitute the raw data to infer the proteins present in the biological source.

Interpretation of the spectral data consists of first matching the fragment ion spectra to their corresponding peptide sequences (peptide spectrum matching) and second to integrate these results to infer the set of proteins initially present in the biological sample (protein inference) [26]. See also **Fig. 1**. These steps are typically automated due to the vast amount of spectra generated in the course of contemporary shotgun proteomics approaches.

Peptide-spectrum matches are typically generated using one of the many available search engines, e.g. [14, 27, 9]. Search engines map fragment ion spectra to the best matching peptide sequence in the protein database of the studied organism [25]. Various statistical measures, such as false discovery rates [4], have been derived to account for possibly incorrect peptide-spectrum matches [6]. In this context the target-decoy strategy has recently grown very popular since it is simple to

implement and compatible with all currently used search engines [23, 13].

Protein inference uses peptide-spectrum matches to infer the identities of proteins initially present in the biological source [26]. A protein identification comprises an assembly of supporting peptide-spectrum matches. Protein inference engines integrate possibly redundant spectral evidence to compile a set of protein identifications that are expected to be correct, i.e. contain at least one correct peptide-spectrum match. The more complex a proteome the more frequently peptide-spectrum matches turn out to ambiguously map to several protein entries, e.g. protein splice variants. It is common practice to circumvent this issue by effectively reducing a protein identification to a gene locus identification in case of ambiguity ("gene locus inference")[1, 5, 3, 34]. More sophisticated protein inference engines though implement statistical or algorithmic approaches to disambiguate peptide-spectrum matches where required [24, 31, 16, 30, 36, 28, 37]. After having applied an inference engine, it is common practice to exclude possibly unreliable protein identifications, such as single hit protein identifications. There has been considerable debate about whether this kind of post-processing enhances protein inference [18, 17].

Protein identifications are not perfect since peptide-spectrum matches can be spurious. Errors at the level of peptide-spectrum matches though propagate non-trivially to the level of protein identifications. While error rate estimation for peptide-spectrum matches is well established [6], several attempts have been made to control protein identification error rates throughout. Besides their inference functionality, most protein inferences engines also estimate error rates based on probabilities of individual peptide-spectrum matches being wrong [24, 31, 16, 28]. It turns out, however, that this kind of approach does not scale well with dataset size [29]. Another approach estimates the number of incorrect protein identifications assuming that false positive peptide-spectrum matches distribute according to a Poisson distribution across the protein database. [1, 35]. The estimates from such models though give ambiguous estimates depending on assumptions regarding single hit protein identifications. Simple target-decoy approaches, estimating the number of false positive protein identifications by the number of decoy identifications [25, 28, 37, 18], have shown to be biased towards too pessimistic estimates [29]. Considering the limitations of the latter approaches,

none of these techniques qualifies as a general purpose method to control protein identification error rates for datasets of different quality and size. To close this gap, we recently proposed the Mayu approach that appropriately adapted the target-decoy strategy to the protein inference task and achieved accurate, independently validated protein identification false discovery rates (i.e. the expected proportion of incorrect among all accepted identifications) for a range of diverse datasets differing in size, underlying proteome and experimental setting [29].

The literature does not provide a starting point to decide which protein inference engine to choose for a particular application scenario (in contrast to the rich literature about search engine comparisons, see e.g.[19]). Specifically, none of the studies presenting a novel protein inference engine [24, 31, 16, 28] reports identification performance by means of a thorough benchmark against at least one baseline method. Instead each study shows that its approach is to some extent able to recover protein identities from different kinds of datasets, i.e. artificial, well characterized protein mixtures comprising at most dozens of sufficiently abundant proteins and real world complex whole proteome mixtures. A single study ([28]) shows results of a competing approach ([24]). This study lacks, however, a systematic benchmark with a sensible performance measure.

Typically performance of a protein inference engine is positively correlated with the number of protein identifications attributed to the respective engine. In this context, protein inference performance is sensibly measured by specifying the number of correct and incorrect protein identifications, i.e. by not only counting the total number of protein identifications but by also considering identification specificity. In the case of an artificial protein mixture, identification performance is easily measured since the protein composition is known and identifications are therefore trivially recognized as true or false positive. In the real world case, identification performance is not straightforward to measure since the true protein composition of the test sample is not known. As delineated before, it is has been only partially understood how to count the number of correct and incorrect protein identifications in this case until recently [29] and therefore protein inference engine performance has only been reasonably approximated and reported for scenarios that do not reflect the heterogeneity and size of contemporary shotgun proteomics datasets [24, 31, 16, 28].

The present study contributes a sensible and generic performance measure that enables to easily benchmark protein inference engines (**Fig. 1**). This measure evaluates the number of true and the proportion of false positives in a particular set of protein identifications. We show that these numbers can be easily and generically estimated on the basis of protein false discovery rates [29] (Supplementary Figure 1). We apply this performance measure to compare a family of widely used protein inference engines. This family is based on the popular "gene locus inference" approach [1, 5, 3, 34] . For this base protein inference engine we additionally study the impact of post-processing schemes related to the exclusion of protein identifications subsets featuring low spectral counts. We report a target-decoy strategy for local false discovery rates [12] to quantify the reliability of various protein identification subsets. In order to systematically study the exclusion of protein identifications after applying one of the base inference engines, we introduce the concept of a selection scheme that formally characterizes properties of a subset of protein identifications (**Fig. 2**). By systematically varying selection schemes we effectively benchmark thousands of different variants of the base protein inference engine (**Fig. 1**). Finally, we apply the benchmark strategy to compare "gene locus inference" with ProteinProphet. For the largest reported shotgun proteomics dataset for *C. elegans* [34] we find that "gene locus inference" without any further pruning achieves the highest performance.

2 Material and Methods

2.1 Spectral data and data processing

This study covers three different data sets from studies varying in MS instrumentation and underlying organism. All studies were based on multidimensional fractionation techniques and comprised samples from *Caenorhabditis elegans* (1,305 LC-MS/MS runs, 5,897,279 fragment ion spectra) [34], *Leptospira interrogans* (24 LC-MS/MS runs, 60518 fragment ion spectra [33]), and *Schizosaccharomyces pombe* (16 LC-MS/MS runs, 40407 fragment ion spectra, manuscript in prep.). The first data set has been acquired on a low resolution LTQ instrument, the latter two were acquired on a high mass accuracy LTQ-FT instrument. The *C. elegans* proteome data are available on Pep-

tideAtlas [10]. We searched each data set against a target-decoy database with Turbo Sequest [14] and Sequest on a Sorcerer machine (Sorcerer^{TM}-SEQUEST®, 3.10.4 release). The search results were processed using the Trans-Proteomic Pipeline [20] to the level of PeptideProphet [21]. The resulting pepXML files were then further analyzed with the Mayu software. If a peptide existed in more than one protein sequence the hit was associated with one protein representing the gene locus or in case of ties to the alphabetically first entry of the protein database ("gene locus identification") [34]. We performed all the database searches using a concatenated target-decoy database [13]. As target database for the *C. elegans* data set we chose wormpep170 (WormBase). For the *L. interrogans* data set we used NC_005824 (National Center for Biotechnology Information), and for the *S. pombe* data set we used 78.S_pombe (European Bioinformatics Institute). As decoy databases we used the reversed sequences of the target database.

2.2 Protein identification false discovery rates with Mayu

We used Mayu to estimate false discovery rates for protein identifications. Mayu is a statistical validation tool that is applied after having performed protein inference, e.g. after having run gene locus inference or ProteinProphet. Specifically, Mayu takes the result of protein inference, i.e. a list of target/decoy protein identifications as input and estimates the false discovery rate of those identifications mapping to the target database. The estimation procedure from the input identification procedures is detailed in [29]. Briefly, Mayu extends the target-decoy approach to estimate false discovery rates for peptide-spectrum matches to protein identifications. Starting off with the list of protein identifications supplied by the protein inference engine, Mayu estimates protein false discovery rates from the counts of identifications mapping to the target- and to the decoy database by means of a hypergeometric model. Instead of simply estimating the expected number of false positive protein identifications as the number of decoy identifications [25, 28, 37, 18], this model additionally corrects for a class of true positive identifications that are supported by both spurious and correct peptide-spectrum matches. The validity of this approach and the accuracy of its false discovery estimates have been confirmed by various statistical and experimental approaches [29]. For a benchmark of Mayu using a defined protein standard mix ([22], see Supplementary Figure 1). The current Mayu implementation is available at

http://tools.proteomecenter.org/wiki/index.php?title=Software:Mayu and takes a list of score annotated target/decoy peptide/protein identifications as input. Mayu supports several data formats, such as pepXML or Mascot csv files. Mayu processes this input to estimate protein false discovery rates for identification sets filtered according to user defined score thresholds. Mayu's main output consists of a csv table reporting several statistics for the identification sets. Amongst other statistics, the total number of protein identifications, their false discovery rate and respective number of expected true positive protein identifications are reported. These statistics are either directly used to evaluate protein inference performance or supplied to evaluate local false discovery rates for pruning strategies (see section **2.3**).

2.3 Local false discovery rates for protein identification subsets

Local false discovery rates can be used to quantify the reliability of subsets of protein identification supplied by a protein inference engine. We use simple properties, such as number of supporting peptide-spectrum matches, to characterize protein identification subsets. More generally, we use an individual property Y to split the complete set of protein identifications into subsets, each featuring the same property value (e.g. single hits) and to measure their quality by local false discovery rates $\mathrm{FDR}(y)$ [12]. By definition of local false discovery rates we can write

$$\mathrm{FDR}(y) \quad = \quad \frac{P(y \mid \mathrm{fp}) \cdot P(\mathrm{fp})}{P(y)} \tag{1}$$

While y corresponds to the property value of a single identification, fp denotes the identification to be false positive. $\mathrm{FDR}(y)$ thus corresponds to $P(\mathrm{fp})$ scaled by the ratio of $P(y \mid \mathrm{fp})$ to $P(y)$. Calculation of $\mathrm{FDR}(y)$ requires to specify the distributions $P(y \mid \mathrm{fp})$, $P(\mathrm{fp})$ and $P(y)$. $P(y)$ can be estimated with its empirical distribution defined by all protein identifications mapping to the target database. Recalling that protein identifications mapping to the decoy database are false positive by definition, $P(y \mid \mathrm{fp})$ can be approximated analogously by its empirical distribution defined by all decoy protein identifications. The protein identification false-discovery rate for the complete identification set is straightforwardly estimated with Mayu (for details please see [29]) and provides an estimate for the prior $P(\mathrm{fp})$, finally allowing to estimate the local false discovery

rate by plugging in the latter estimates.

2.4 Protein identification selection schemes

Pruning unreliable (or selecting reliable) protein identifications after having run a protein inference engine is a popular postprocessing procedure to enrich for correct identifications. We use selection schemes to characterize (presumably high quality) protein identification subsets that we wish to report in the final identification list. A very simple selection scheme could for instance characterize the subset of all non-single hit protein identifications and the single-hit identifications whose peptide-spectrum matches score higher than any decoy match.

Generally, we introduce the notion of a selection scheme to formally express postprocessing identifcation selection rules of the form: accept single hits with PSM FDR lower than 0.1%, accept double hits with PSM FDR lower than 0.2%, accept triple hits with PSM FDR lower than 1%, ... , where "with PSM FDR" is shorthand for "supported by peptide-spectrum matches with false discovery rate". Formally, a selection scheme is characterized by a sequence of peptide-spectrum match false discovery rate $m_1, m_2, m_3, ...$ The selection scheme considers those protein identifications which are supported by at least i peptide-spectrum matches that map to the peptide-spectrum match set with false discovery rate less than m_i. In the forgoing example we would have $m_1 = 0.1\%, m_2 = 0.2\%, m_3 = 1\%$. Selection schemes thus allow us to define protein identification sets where protein identifications evidenced by very few high confidence peptide-spectrum matches and protein identifications supported by a large number of lower confidence peptide-spectrum matches. For an illustration see also **Fig. 2**.

2.5 Screening and false discovery rate evaluation of selection schemes

For our selection scheme screening we consider false discovery rate thresholds for single-, double-, triple-, quadruple-, quintuple- and sextuple hits or higher order protein identifications. We exhaustively enumerate all selection schemes for the respective false discovery thresholds $m_1, ..., m_6, m_{>6}$ allowing each to assume a value from a panel of twelve thresholds (0%, 0.01%, 0.02%, 0.03%, 0.04%, 0.05%, 0.1%, 0.15%, 0.2%, 0.25%, 0.3% or 0.35%). Considering that a sensible selection

scheme filters protein identifications with less spectral support more stringently ($m_i \leq m_j, i \leq j$), this screening considers 12376 different selection schemes. The protein identification false discovery rate for a set of protein identifications obtained by a specific selection scheme is determined as their local false discovery rate estimate (see section **2.3**). This protein identification false discovery rate estimate straightforwardly translates into an estimate for the expected number of true positive protein identifications since the total number of identifications is known. We consider a selection scheme optimal with respect to a chosen protein identification false discovery rate if its respective protein identification set maximizes the expected number of *true positive* protein identifications.

2.6 Protein inference engine performance & benchmark

For the protein inference engine benchmark we assume that for each competitor the (target-decoy) identification results are available in terms of an identification list. We measure performance of each competing approach by evaluating the absolute number of *true positive* identifications and the respective proportion of false positives, i.e. the false discovery rate. Since the total number of identifications is trivially given by the length of a protein identification list it is sufficient to estimate the false discovery rate in order to complete the performance measure. We estimate the false discovery rate directly with Mayu [29]. In case the competitors inference strategy involves a pruning step according to a selection scheme, the proportion of false positives is estimated as local false discovery rate as described in the preceding section.

Most inference engines assign a score to each protein identification and therefore produce a series of protein identifications with increasing size and false discovery rate. Assessing this series of identification sets yields a response curve that characterizes the performance of the inference engine across the whole spectrum of false discovery rates. In summary, competitors can now be sensibly ranked according to the amount of *true positive* identifications at a user defined false discovery rate. See also **Fig. 1**.

3 Results

In the following we investigate the quantitative impact of certain properties on protein identifications reliability, such as sequence length and spectral support of protein identifications. We go on

and demonstrate the application of our generic protein inference performance measure to systematically benchmark "gene locus inference" in conjunction with a multitude of selection schemes based on spectral support. We conclude by extending the benchmark by the more sophisticated protein inference engine ProteinProphet.

3.1 Local false discovery rates for protein identification subsets

To date, deciding on the final set of protein identifications in a given proteomics dataset is frequently based on heuristic criteria supposed to enrich for valid identifications. A widely used strategy filters for protein identifications whose peptide-spectrum matches score above a certain threshold. More stringent strategies furthermore require a valid protein identification to be composed of a minimal number of supporting peptide-spectrum matches (e.g., neglect all single hits) . There have been substantial debates about the validity of such criteria to compile protein identification sets from large datasets.

To determine whether removal of particular subsets of protein identifications improves the quality of the remaining identifications, we estimated local false discovery rates for identification subsets. These subsets were characterized by a certain property (e.g. supported by single hits) expected to have an impact on the quality of protein identifications (**Fig. 3**). We studied the effect of three properties that were expected to have an impact on the quality of protein identifications. Specifically, we explored the effects of protein sequence length in terms of amino acids, number of peptide-spectrum matches supporting the protein identification (number of supporting peptide-spectrum matches) and peptide-spectrum match alignment type.

In a first step we generated local false discovery rate estimates for protein identification sets characterized by protein sequence length (**Fig. 3a-c**). Increasing protein sequence length is expected to amplify false discovery rate since false positive peptide-spectrum matches map to the database by chance and therefore are more likely to map to larger proteins. Our local false discovery rate estimates clearly confirm this expectation. The local false discovery rate for proteins of for instance sequence length 400 is two fold higher than for proteins with sequence length 100 using a peptide-

spectrum matches false discovery rate cutoff of 0.01. We find a similar though not so pronounced trend for smaller datasets, i.e. subsets of the complete *C. elegans* dataset.

We went on to study protein identification sets characterized by a varying number of supporting peptide-spectrum matches (**Fig. 3d-f**). Note that a protein identification with one supporting peptide can be supported by several peptide-spectrum matches. As expected, we found that the confidence in a protein identification scales with the number of peptide-spectrum matches supporting it. We observed impressively high false discovery rate for protein identifications only supported by a single peptide-spectrum match (single hits) in the complete dataset, exceeding 0.65 for a presumably stringent peptide-spectrum match false discovery rate of 0.01. Even protein identifications being apparently approved by two peptide-spectrum matches complying with a cutoff of 0.01 featured false discovery rate of 0.4. While not being so pronounced, we encounter a similar situation for smaller subsets of the *C. elegans* datasets. The discrepancy between false discovery rate of peptide-spectrum matches and corresponding protein identifications was most pronounced for the subset of single hits. These results confirm a similar discrepancy estimated by other (not so generically applicable) methods, such as validation with synthetic peptides [29], considering deviations in measured and predicted isoelectric point [29] or manual curation [17].

In order to elucidate the influence of the peptide-spectrum match distribution along the sequence of a given protein identification, we categorize protein identifications according to the peptide-spectrum match alignment type score PAT. PAT splits the set of protein identifications into those being composed of either one, two, three or more peptide-spectrum matches. Subsets of identifications containing two or three identifications are further split into cases where peptide-spectrum matches either redundantly map to the same peptide or not. PAT can thus adopt seven different values, each representing a different arrangement of peptide-spectrum matches along a protein sequence (**Fig. 3g-i**). Besides observing the trend towards lower false discovery rates for larger spectral support, it is particularly interesting to compare false discovery rates for different PAT scores within the set of protein identifications composed of either two or three peptide-spectrum matches. For combinatorial reasons, it is to be expected that the least redundant alignment of

peptide-spectrum matches is most likely to happen by chance and therefore to yield the largest false discovery rate. The false discovery rate estimates confirm this expectation for less stringent peptide-spectrum match cutoffs (> 0.006). Surprisingly, for more conservative cutoffs we observe a transition to the opposite. In order to explain this trend, it is useful to distinguish two types of false positive peptide-spectrum matches, those that originate from frequently occurring entities not being present in target or decoy database (e.g. unconsidered modified peptides, chemical impurities) and others whose sources are less frequent entities. We assume that the former group contributes to high scoring false positive peptide-spectrum matches to a larger extent than the latter. Since the same entity is supposed to give rise to similar spectra for each occurrence, it reproducibly maps falsely to the same peptide sequence, consequently making sense to the overrepresentation of redundant alignment types for the more stringently selected peptide-spectrum matches.

In summary, we found that the investigated protein identification properties are powerful indicators of protein identification quality and therefore represent a promising starting point to selectively prune protein identification sets in large to very large datasets to enrich for valid protein identifications.

3.2 Pruning protein identifications does not enhance protein inference

The foregoing analysis suggests that consequently excluding single hits might enrich for correct protein identifications (**Fig. 4a**, w/o single hits). However, it might be more beneficial to selectively retain high quality (very low false discovery rate) single hits (**Fig. 4a**, w/o uncertain single hits). In order systematically assess such kind of hypotheses, we introduced the notion of a selection scheme. A selection scheme formally characterizes which protein identifications obtained from a protein inference engine to select as an entry reported in the final list of identifications (for details see **section 2.4, Fig. 2**). We constructed a multitude of (possibly more involved) selection schemes that retain protein identifications with varying spectral quality depending on the number of supporting peptide-spectrum matches. We demonstrated how to assess and compare these selection schemes by means of our generic protein inference performance measure.

We systematically searched for optimal selection schemes, i.e. selection schemes that maximized the number of expected *true positive* protein identifications for a desired false discovery rate. On the basis of the result of "gene locus inference", 12376 different selection schemes were analyzed in the course of this benchmark. Regardless of the desired protein identification false discovery rate, optimal selection schemes turned out to be the ones that consider all high quality peptide-spectrum matches, irrespective of the "number of supporting peptide-spectrum matches" property (**Fig. 4a**, optimal). For instance, the optimal protein identification set featuring protein false discovery rate of 0.015 is compiled from the complete set of peptide-spectrum matches with false discovery rate of 0.0005. All other selection schemes were inferior. In particular our results clearly ruled out selection schemes that consequently neglected single hits, as well as selection schemes that selectively included protein identifications supported by a large number of low confidence peptide-spectrum matches (**Fig. 4b**).

The forgoing result is confirmed for other datasets. We performed the same analysis for two other datasets from multidimensional fractionation experiments acquired for the bacterium *L. interrogans* (24 LC-MS/MS runs, **Fig. 5a**) and for the eukaryote *S. pombe* (16 LC-MS/MS runs, **Fig. 5b**). Selection schemes considering all spectral evidence up to a certain quality achieved optimal performance according to the before introduced measure. Instead, selection schemes consequently neglecting single hits exhibited significantly lowered performance.

In summary, selection schemes considering all peptide-spectrum matches, including those giving rise to the less reliable single hits, turned out to be optimal for diverse datasets. However, peptide-spectrum matches have to be selected much more carefully than appreciated so far, in order to achieve reasonable protein identification false discovery rate for datasets of large size.

3.3 Simple protein inference engines show competitive performance

We compared "gene locus inference" and its selection scheme variants to ProteinProphet on the datasets for *C. elegans* (**Fig. 4a**), *L. interrogans* (**Fig. 5a**) and *S. pombe* (**Fig. 5b**).

In order to ensure a fair comparison we analyzed the degree of parsimony for "gene locus inference" and ProteinProphet (see also Discussion). For each inference engine we computed a parsimony score, i.e. the ratio of the number of actually reported protein identifications and the lowest possible number of protein identifications explaining all identified peptides (see also Supplementary Material). The results showed that for all datasets and over the range of considered protein identification false discovery rates parsimony scores were practically identical and consistently evaluating around 1.02 for both "gene locus inference" and ProteinProphet. In summary, both approaches achieve the same almost maximal degree of parsimony and can therefore sensibly compared as described above.

We observed that "gene locus inference" without any pruning performs clearly better over the complete range of reasonable protein identification false discovery rates for the larger *C. elegans* dataset. In this situation, ProteinProphet is also inferior to the single hit exclusion schemes for small false discovery rates, though outperforms these for less stringent identification quality requirements. For the comparably smaller datasets for the other organisms ProteinProphet though achieves optimal performance (according to our protein FDR based measure).

At a first glance it might be surprising that a simple protein inference strategy like "gene locus inference" (see also Methods) outperforms a sophisticated inference engine like ProteinProphet when confronted with a large dataset. Considering that ProteinProphet effectively implements a probabilistically motivated selection scheme this result is though consistent with the forgoing analysis. To see this, consider ProteinProphet's probabilistic model that is supposed to also recover proteins that are redundantly evidenced by less reliable fragment ion spectra and to penalize single hit wonders stronger the more multiple hit identifications are present. This situation is particularly extreme in a large dataset like the one studied here. In the previous analysis, we systematically assessed all reasonable selection schemes for the complete *C. elegans* dataset and proved them all to be inferior to the simple "gene locus inference" without any further pruning. These results suggest that, at least for large datasets like the *C. elegans* dataset, low quality spectra do not contribute novel protein identifications and potentially mislead approaches that aim to exploit them as an

additional information source.

4 Discussion

This work systematically assesses how pruning unreliable protein identification subsets affects protein inference performance. An exploratory study aimed at investigating possibly unreliable protein identifications subsets. We assessed reliability by means of local false discovery rates. We further studied whether pruning such unreliable protein identifications is beneficial for protein inference performance. Therefore we first introduced the concept of selection schemes to enable a systematic enumeration of thousands of conceivable pruning strategies (**Fig. 2**, Methods). In a second step we introduced a performance measure that allowed to generically compare the protein inference results obtained by each of the many pruning strategies. This measure evaluates the number of *correct* identifications and the involving false discovery rate. We applied this measure to benchmark pruning strategies defined by selection schemes computed for protein identifications obtained by the "gene locus inference" approach on a variety of shotgun proteomics datasets.

We studied the influence of various protein identification properties on identification false discovery rates. We observed that the number of peptide-spectrum matches supporting a protein identification has a severe impact on the identifications reliability. We therefore focused our subsequent pruning strategies on this property. The generic concept of selection schemes though also lends itself to define pruning strategies based on other protein identification properties, such as protein length. Future studies might benefit from incorporating these, too.

Our results confirm a recently appeared study that advocates to retain single hit wonders instead of discarding them [18] since these typically comprise many correct identifications. Here we studied these two pruning strategies among thousands of other strategies defined by the selection schemes. We could consistently rule out all conceivable pruning strategies to improve protein inference performance. While our heterogeneous dataset is presumably representative of most large scale shotgun proteomics datasets, it is still conceivable that for other datasets these conclusions do not hold. Consider for instance repetitive measurements of a large number of very similar samples. Such

a scenario might result in a situation where true single hits become exceedingly rare, redeeming selection schemes that exclude single hits.

It is furthermore to be expected that the results of a pruning strategy benchmark differ with the underlying protein inference engine. Our results for pruning strategies for "gene locus inference", in particular regarding inclusion of single hits, do not necessarily carry over to other protein inference engines. One reason for different behaviors across inference engines are hidden internal procedures to produce the final list of identifications. These frequently preclude to trace back the spectral evidence supporting the individual protein identifications and therefore blur the concept of a one-hit-wonder. To illustrate this point, imagine an (extreme) protein inference engine that internally neglects all ambiguous peptide-spectrum matches (since it deems them confusing and misleading) and in addition only reports the highest scoring match per protein. This inference engine would solely report one-hit-wonders. It would though not be surprising to find these one-hit-wonders to achieve reasonable protein FDR. In conclusion, we advocate to judge the usefulness of pruning strategies anew for each combination of dataset and protein inference engine. This decision can be made straightforwardly by means of the benchmark framework described and exemplarily demonstrated in this study.

Our performance measure can be straightforwardly applied to benchmark a user defined set of protein inference engine candidates. In this study we benchmark ProteinProphet and "gene locus inference" including its selection scheme variants. The approach is, however, equally applicable to a wide range of inference engines since the evaluation of the performance criterion is performed on the list of (target-decoy) protein identifications. It is not strictly necessary to compare the same protein inference engines as in the presented study. This study focused on many pruning variants of a simple protein inference engine to systematically explore the impact of excluding specific protein identification subsets. For many typical application scenarios the user instead relies on a small set of (e.g. three) well established protein inference engines. This small set of inference engines can equally well be the basis of such a benchmark. After having processed the mass spectrometrical data with each of the considered inference engines, having filtered the protein identifications to

achieve a user defined protein false discovery rate [29], their performance in terms of number of true protein discoveries can be easily evaluated, see also **Fig. 1**. Depending on the application scenario, the user might allow for a lower or larger number of false positive protein identifications (lower or larger protein false discovery rate) to assess the performance of the competing inference engines. By this means our approach enables the experimentalist to perform a benchmark and to make a choice that are tailored to his application.

In order to ensure a fair comparison, competing protein inference engines should be comparable with respect to our performance measure. Since our performance measure rewards large numbers of *correct* protein identifications, the competitors should base their inference on a similar sized repertoire of possible protein identities. This requirement is mainly ensured by providing the same protein database to all competitors. However, it is conceivable that cases arise, where a protein inference engine would be intrinsically disadvantaged if it were to infer less resolved entities, such as exclusively gene loci, compared to competitors that could possibly report a larger number of higher resolved entities, such as splice variants[1]. Furthermore it is necessary to ensure that the compared inference engines map the underlying peptide identities to the repertoire of possible protein identities with a similar degree of parsimony. This requirement is related to the general difficulty (for any statistical validation approach) to estimate the frequency of erroneous protein identifications that stem from correct peptide identifications associated to the wrong protein. Comparing protein inference engines of varying degree of parsimony would therefore favor more generous approaches that would generate more undetected spurious protein identifications. Most protein inference engines effectively achieve sets of protein identifications that are either actually or at least practically maximally parsimonious and thereby of equivalent degree of parsimony. We have specifically shown that both approaches compared in this study, i.e. "gene locus inference" and ProteinProphet achieve practically maximally parsimonious protein identifications sets from the various considered datasets and can therefore be sensibly compared by our benchmark approach.

For the protein inference engine benchmark, we optimize a tradeoff between identification sensi-

[1]Note that this example does not apply to "gene locus inference" as introduced here. Given unambiguous peptide-spectrum matches, "gene locus inference" is well able to identify splice variants.

tivity and specificity. Despite being appealing, this objective is not necessarily always suitable. This is particularly the case where shotgun proteomics studies focus on a small set of proteins and aim to make explicit or even resolve possible ambiguities, such as gene products of a single gene locus. These cases necessitate protein inference engines like ProteinProphet [24] or IDPicker [37] that provide a protein grouping functionality.

Our results have implications on guidelines for reporting lists of protein identifications. Such guidelines are typically designed to ensure the reliability of protein identifications. Currently, rigid rules, like general exclusion of single hit wonders, are discussed to be considered for these guidelines. Our study shows that for our datasets, lists of protein identifications excluding single hit wonders are indeed very reliable, i.e. feature very low false discovery rates, though ignore evidence for many true positive hits. Furthermore, our benchmark revealed that, for the studied datasets, accounting for all spectral evidence of sufficient quality is the preferable inference strategy, since this approach achieved the same reliability in terms of false discovery rates at a significantly larger number of *true positive* protein identifications. Although this result is confirmed for diverse datasets, it is still conceivable that other datasets might behave differently and therefore benefit from another inference strategy. Our study exemplifies that reporting guidelines should not require protein identifications to comply with rigid rules ruling out possibly superior protein inference strategies. Instead the reporting guidelines should leave the choice of the protein inference strategy to the researcher and only demand the resulting identifications to meet an objective reliability measure, such as false discovery rate.

In the context of large shotgun proteomics projects aiming at extensive proteome coverage it is desirable to (1) decide upon the experiments that are expected to produce the most informative data, i.e. to most effectively explore a proteome [15, 5, 7, 8] and (2) to optimally evaluate the finally acquired data, i.e. to optimally perform protein inference. The presented benchmark approach contributes to the second step by generically enabling to choose the best protein inference engine among those under consideration for a specific shotgun proteomics dataset under study. We apply this approach to benchmark thousands of pruning strategies for protein inference for diverse shotgun

proteomics datasets. For all considered cases, processing of the data with simple protein inference approaches and keeping all the spectral evidence achieves competitive proteome coverage.

Acknowledgments

We thank Sabine Schrimpf for providing the *C. elegans* dataset and Alexander Schmidt for contributing the *L. interrogans* and *S. pombe* datasets. We acknowledge SystemsX.ch and the Center for Systems Physiology and Metabolic Diseases for funding.

References

[1] M. Adamski, T. Blackwell, R. Menon, L. Martens, H. Hermjakob, C. Taylor, and GS Omenn. Data management and preliminary data analysis in the pilot phase of the HUPO Plasma Proteome Project. *Proteomics*, 5(13):3246, 2005.

[2] R. Aebersold and M. Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207, 2003.

[3] Katja Baerenfaller, Jonas Grossmann, Monica A. Grobei, Roger Hull, Matthias Hirsch-Hoffmann, Shaul Yalovsky, Philip Zimmermann, Ueli Grossniklaus, Wilhelm Gruissem, and Sacha Baginsky. Genome-Scale Proteomics Reveals Arabidopsis thaliana Gene Models and Proteome Dynamics. *Science*, 320(5878):938–941, 2008.

[4] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[5] E. Brunner, C. H. Ahrens, S. Mohanty, H. Baetschmann, S. Loevenich, F. Potthast, E. W. Deutsch, C. Panse, U. de Lichtenberg, O. Rinner, H. Lee, P. G. Pedrioli, J. Malmstrom, K. Koehler, S. Schrimpf, J. Krijgsveld, F. Kregenow, A. J. Heck, E. Hafen, R. Schlapbach, and R. Aebersold. A high-quality catalog of the Drosophila melanogaster proteome. *Nat Biotechnol*, 25(5):576–83, 2007.

[6] H. Choi and A.I. Nesvizhskii. False discovery rates and related statistical concepts in mass spectrometry-based proteomics. *Journal of Proteome Research*, 7(01):47–50, 2007.

[7] M. Claassen, R. Aebersold, and J. M. Buhmann. Proteome coverage prediction with infinite Markov models. *Bioinformatics*, 25(12):i154–60, 2009.

[8] M. Claassen, R. Aebersold, and J. M. Buhmann. Proteome Coverage Prediction for Integrated Proteomics Datasets. *Journal of Computational Biology*, 18(3):283–29, 2011.

[9] R. Craig and R.C. Beavis. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, 20(9):1466–1467, 2004.

[10] F. Desiere, E. W. Deutsch, A. I. Nesvizhskii, P. Mallick, N. L. King, J. K. Eng, A. Aderem, R. Boyle, E. Brunner, S. Donohoe, N. Fausto, E. Hafen, L. Hood, M. G. Katze, K. A. Kennedy, F. Kregenow, H. Lee, B. Lin, D. Martin, J. A. Ranish, D. J. Rawlings, L. E. Samelson, Y. Shiio, J. D. Watts, B. Wollscheid, M. E. Wright, W. Yan, L. Yang, E. C. Yi, H. Zhang, and R. Aebersold. Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol*, 6(1):R9, 2005.

[11] B. Domon and R. Aebersold. Options and considerations when selecting a quantitative proteomics strategy. *Nat Biotechnol*, 28(7):710–21, 2010.

[12] B. Efron and R. Tibshirani. Empirical bayes methods and false discovery rates for microarrays. *Genet Epidemiol*, 23(1):70–86, 2002.

[13] J. E. Elias and S. P. Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods*, 4(3):207–14, 2007.

[14] J.K. Eng, A.L. McCormack, J.R. Yates, et al. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, 5(11):976–989, 1994.

[15] J. Eriksson and D. Fenyo. Improving the success rate of proteome analysis by modeling protein-abundance distributions and experimental designs. *Nat Biotechnol*, 25(6):651–5, 2007.

[16] J. Eriksson, D. Fenyo, et al. Probity: a protein identification algorithm with accurate assignment of the statistical significance of the results. *Journal of Proteome Research*, 3(1):32–36, 2004.

[17] Monica A. Grobei, Ermir Qeli, Erich Brunner, Hubert Rehrauer, Runxuan Zhang, Bernd Roschitzki, Konrad Basler, Christian H. Ahrens, and Ueli Grossniklaus. Deterministic protein inference for shotgun proteomics data provides new insights into Arabidopsis pollen development and function. *Genome Research*, 19(10):1786–1800, 2009.

[18] N. Gupta and P. Pevzner. False discovery rates of protein identifications: a strike against the two-peptide rule. *Journal of Proteome Research*, 8(9):4173–4181, 2009.

[19] EA Kapp, F. Schütz, LM Connolly, JA Chakel, JE Meza, CA Miller, D. Fenyo, JK Eng, JN Adkins, GS Omenn, et al. An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis. *Proteomics*, 5(13):3475, 2005.

[20] A. Keller, J. Eng, N. Zhang, X. Li, and R. Aebersold. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Molecular systems biology*, 1, 2005.

[21] A. Keller, A. I. Nesvizhskii, E. Kolker, and R. Aebersold. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem*, 74(20):5383–92, 2002.

[22] John Klimek, James S. Eddes, Laura Hohmann, Jennifer Jackson, Amelia Peterson, Simon Letarte, Philip R. Gafken, Jonathan E Katz, Parag Mallick, Hookeun Lee, Alexander Schmidt, Reto Ossola, Jimmy K. Eng, Ruedi Aebersold, and Daniel B Martin. The standard protein mix database: A diverse data set to assist in the production of improved peptide and protein identification software tools. *Journal of Proteome Research*, 7(1):96–103, 2008.

[23] R.E. Moore, M.K. Young, and T.D. Lee. Qscore: an algorithm for evaluating SEQUEST database search results. *Journal of the American Society for Mass Spectrometry*, 13(4):378–386, 2002.

[24] A. I. Nesvizhskii, A. Keller, E. Kolker, and R. Aebersold. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem*, 75(17):4646–58, 2003.

[25] A. I. Nesvizhskii, O. Vitek, and R. Aebersold. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat Methods*, 4(10):787–97, 2007.

[26] A.I. Nesvizhskii and R. Aebersold. Interpretation of shotgun proteomic data: the protein inference problem. *Molecular & Cellular Proteomics*, 4(10):1419, 2005.

[27] D.N. Perkins, D.J.C. Pappin, D.M. Creasy, J.S. Cottrell, et al. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–3567, 1999.

[28] Thomas S. Price, Margaret B. Lucitt, Weichen Wu, David J. Austin, Angel Pizarro, Anastasia K. Yocum, Ian A. Blair, Garret A. FitzGerald, and Tilo Grosser. EBP, a Program for Protein Identification Using Multiple Tandem Mass Spectrometry Datasets. *Mol Cell Proteomics*, 6(3):527–536, 2007.

[29] Lukas Reiter, Manfred Claassen, Sabine P. Schrimpf, Marko Jovanovic, Alexander Schmidt, Joachim M. Buhmann, Michael O. Hengartner, and Ruedi Aebersold. Protein Identification False Discovery Rates for Very Large Proteomics Data Sets Generated by Tandem Mass Spectrometry. *Mol Cell Proteomics*, 8(11):2405–2417, 2009.

[30] Katheryn A. Resing, Karen Meyer-Arendt, Alex M. Mendoza, Lauren D. Aveline-Wolf, Karen R. Jonscher, Kevin G. Pierce, William M. Old, Hiu T. Cheung, Steven Russell, Joy L. Wattawa, Geoff R. Goehle, Robin D. Knight, and Natalie G. Ahn. Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. *Analytical Chemistry*, 76(13):3556–3568, 2004.

[31] R.G. Sadygov, H. Liu, and J.R. Yates. Statistical models for protein validation using tandem mass spectral data and protein amino acid sequence databases. *Anal. Chem*, 76(6):1664–1671, 2004.

[32] A. Schmidt, N. Gehlenborg, B. Bodenmiller, L. N. Mueller, D. Campbell, M. Mueller, R. Aebersold, and B. Domon. An integrated, directed mass spectrometric approach for in-depth characterization of complex peptide mixtures. *Mol Cell Proteomics*, 7(11):2138, 2008.

[33] Alexander Schmidt, Martin Beck, Johan Malmstrom, Henry Lam, Manfred Claassen, David Campbell, and Ruedi Aebersold. Absolute quantification of microbial proteomes at different states by directed mass spectrometry. *Mol Syst Biol*, 7, 2011.

[34] S.P. Schrimpf, M. Weiss, L. Reiter, C.H. Ahrens, M. Jovanovic, J. Malmström, E. Brunner, S. Mohanty, M.J. Lercher, P.E. Hunziker, et al. Comparative functional analysis of the Caenorhabditis elegans and Drosophila melanogaster proteomes. *PLoS Biol*, 7(3):e48, 2009.

[35] D.J. States, G.S. Omenn, T.W. Blackwell, D. Fermin, J. Eng, D.W. Speicher, and S.M. Hanash. Challenges in deriving high-confidence protein identifications from data gathered by a HUPO plasma proteome collaborative study. *Nature Biotechnology*, 24(3):333, 2006.

[36] Xiaoyu Yang, Vijay Dondeti, Rebecca Dezube, Dawn M. Maynard, Lewis Y. Geer, Jonathan Epstein, Xiongfong Chen, Sanford P. Markey, and Jeffrey A. Kowalak. Dbparser: web-based software for shotgun proteomic data analyses. *Journal of Proteome Research*, 3(5):1002–1008, 2004.

[37] B. Zhang, M.C. Chambers, D.L. Tabb, et al. Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. *J. Proteome Res*, 6(9):3549–3557, 2007.

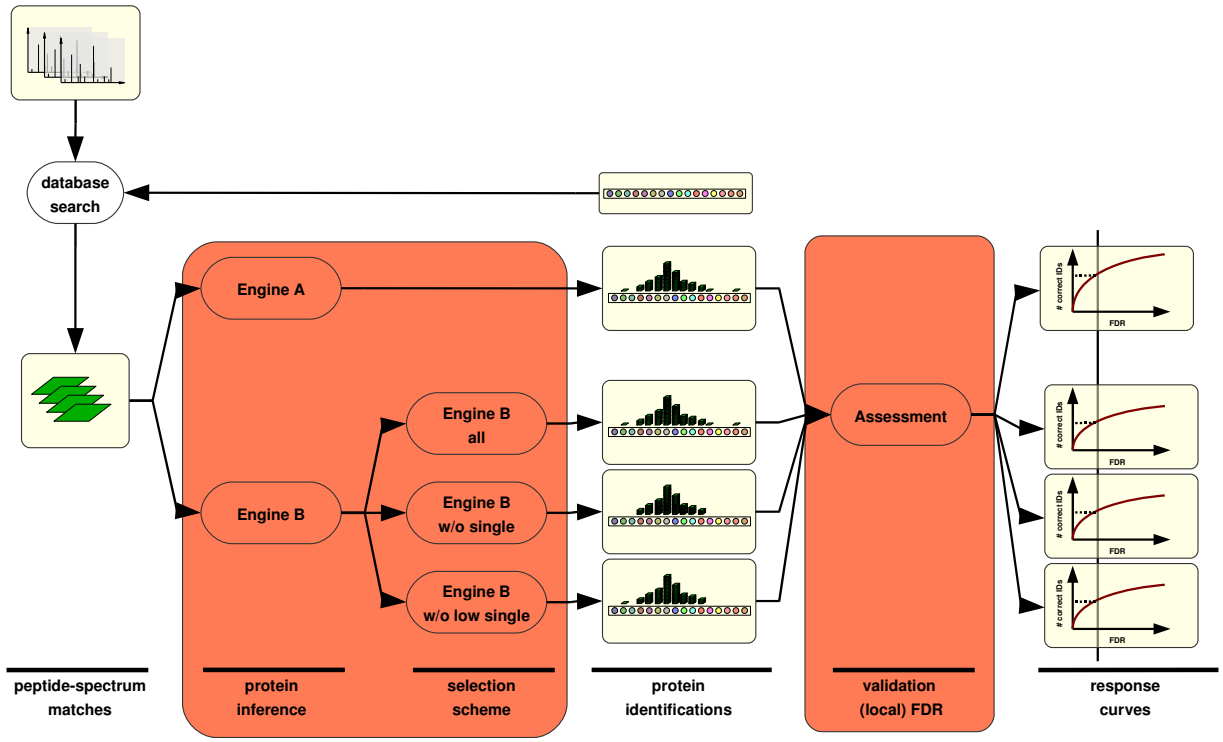

Figure 1: Protein inference engine benchmark schema. Tandem mass spectra are generated in the course of a shotgun proteomics experiment. Protein identities are recovered in two distinct steps, i.e. (1) peptide identification yielding peptide-spectrum matches and (2) protein inference assembling peptide-spectrum matches to protein identifications. Optionally, protein inference is followed by additionally pruning particular protein identifications sets, e.g. single hit identifications. We formally characterize these sets by means of selection schemes to systematically study different pruning strategies. Protein identification reliability is assessed in terms of (possibly local) protein identification false discovery rates. Protein inference performance is measured by estimating the number of *correct* identifications over a range of different protein identification false discovery rates, thereby giving rise to inference engine characteristic response curves. Comparison and ranking of protein inference engines is usually performed for a user defined protein identification false discovery rate. Processes studied in this work are highlighted in red. Specifically, these are (1) selection scheme variants of available protein inference engines and (2) assessment and comparison of protein inference performance.

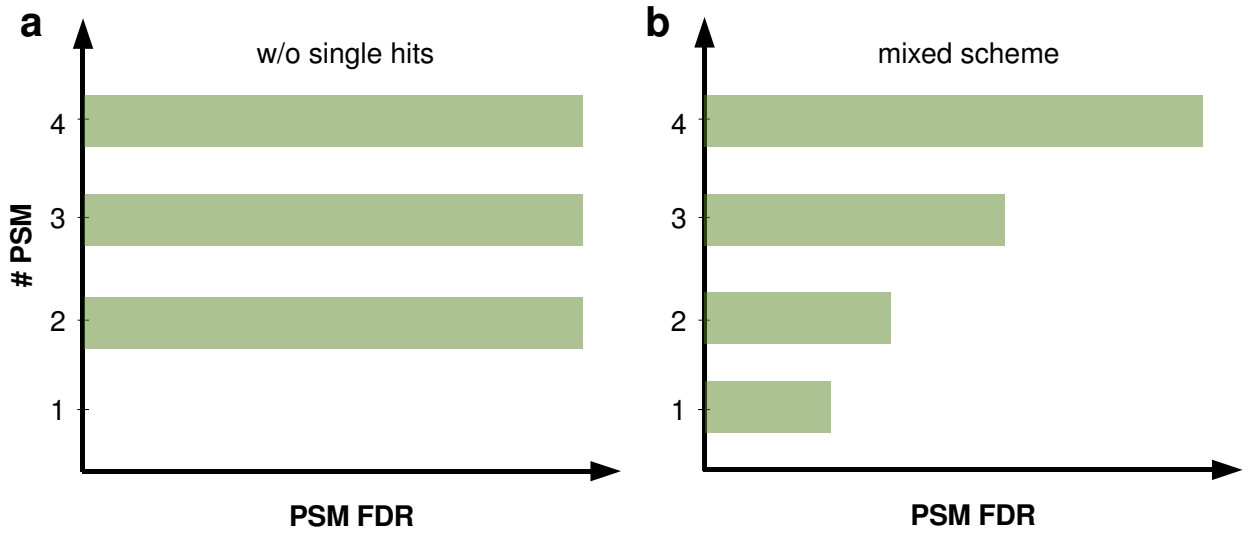

Figure 2: Selection scheme illustration. Selection schemes aim to formalize the notion of selecting the spectra more stringently for protein identifications evidenced by few spectra than for those featuring more redundant evidence. Selection schemes characterize protein identification subsets according to the reliability of peptide spectrum matches (PSM FDR) and some property of a protein identification, e.g. the number of supporting peptide spectrum matches (# PSM). Formally, a selection scheme specifies a series of peptide-spectrum match false discovery thresholds $m_1, m_2, ...$ and accordingly considers protein identifications that for some $i = 1, 2, 3, ...$ are supported by at least i peptide-spectrum matches afflicted with false discovery rate of less than m_i. **(a)** depicts the selection scheme for excluding all single hit protein identifications and considering all other protein identifications supported by at least two peptide-spectrum matches at false discovery rate lower than some threshold. **(b)** depicts a more intricate selection scheme that allows to consider single hit protein identifications as long as the respective peptide-spectrum matches feature a low false discovery rate. With increasing support the spectral quality requirements decrease.

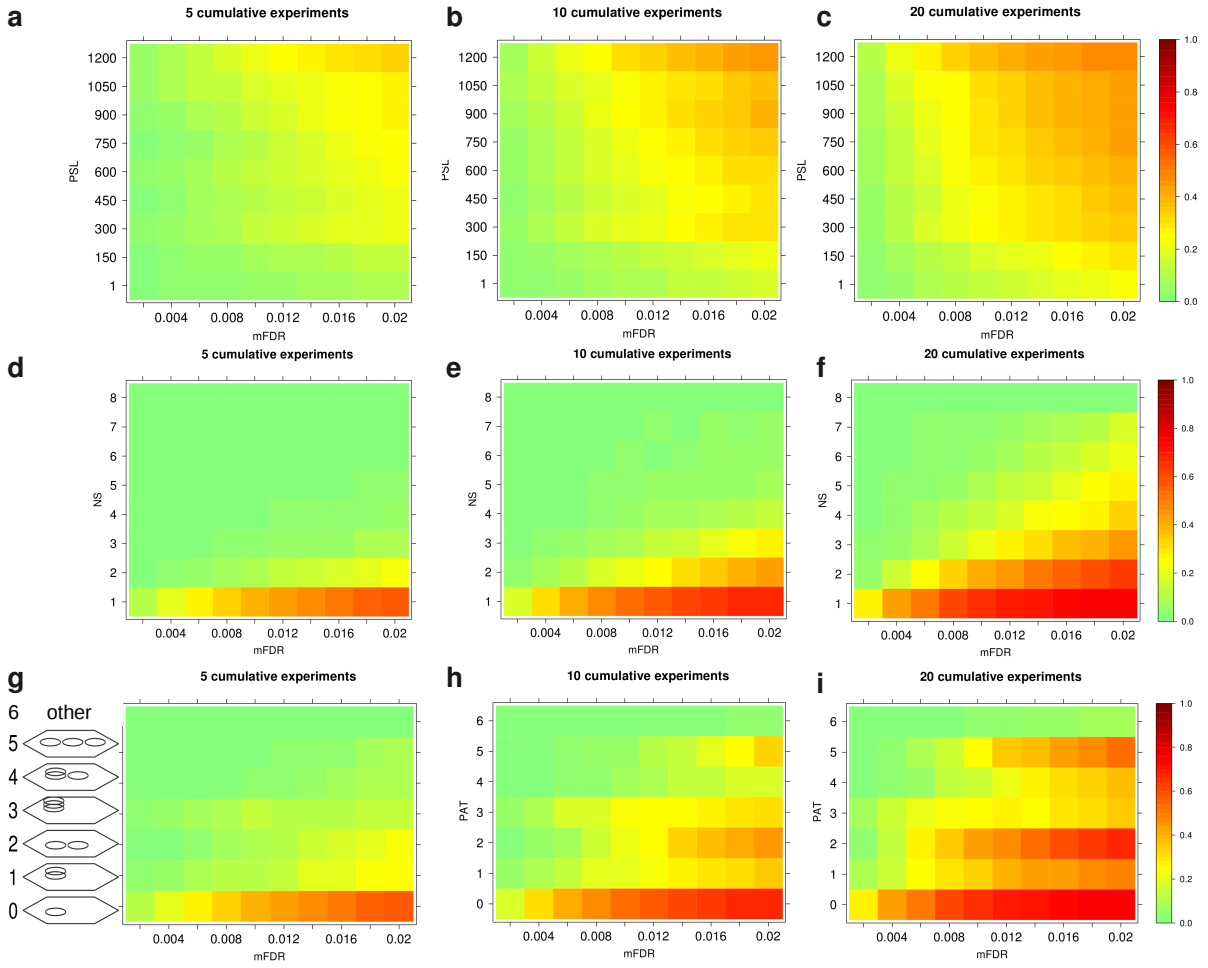

Figure 3: Local protein identification false discovery rate (FDR) for protein identification groups characterized by either protein sequence length in amino acids PSL (specified by lower bin boundaries) (a-c), number of peptide-spectrum matches defining the identification NS (d-f) or peptide-spectrum match alignment type PAT (g-i). Each heat map depicts results for dataset partitions of varying size (amount of cumulative experiments). Magnitude of local protein identification false discovery rate is color coded as indicated. Certain protein identification subsets feature more than 60 fold higher protein identification false discovery rate than the underlying peptide-spectrum match false discovery rate (mFDR).

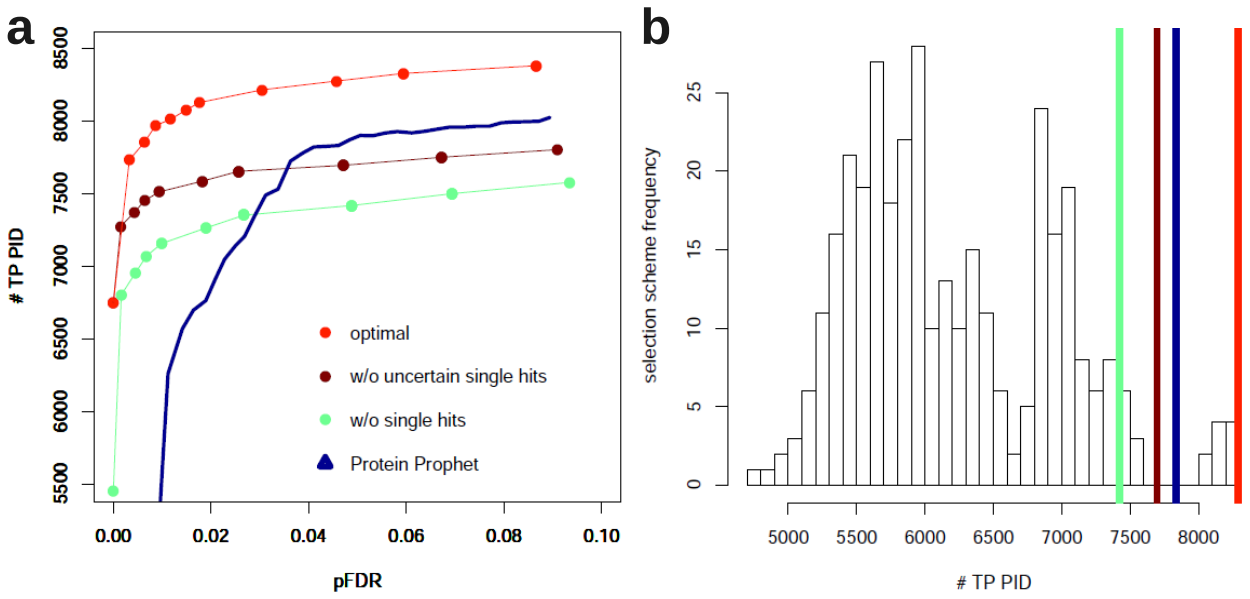

Figure 4: Optimal pruning strategies with respect to expected number of *true positive* protein identifications (# TP PID) at varying protein identification false discovery rate (pFDR) levels. **(a)** Comparison of protein identification selection schemes. All optimal selection schemes (see text) consider all peptide-spectrum matches (PSMs) of sufficiently low peptide-spectrum match false discovery rate (red). Two alternative selection schemes are shown exemplarily. Selection schemes consequently neglecting single hits (green) or solely neglecting single hits with nonzero uncertainty (brown). The response curve of ProteinProphet is shown in blue.**(b)** Histogram of expected number of *correct* protein identifications for all selection schemes at protein identification false discovery rate < 0.05. The performance of the exemplary schemes and ProteinProphet are plotted according to their color code in (a). While three groups can be discerned, the clearly detached top group only considers selection schemes retaining single hit identifications.

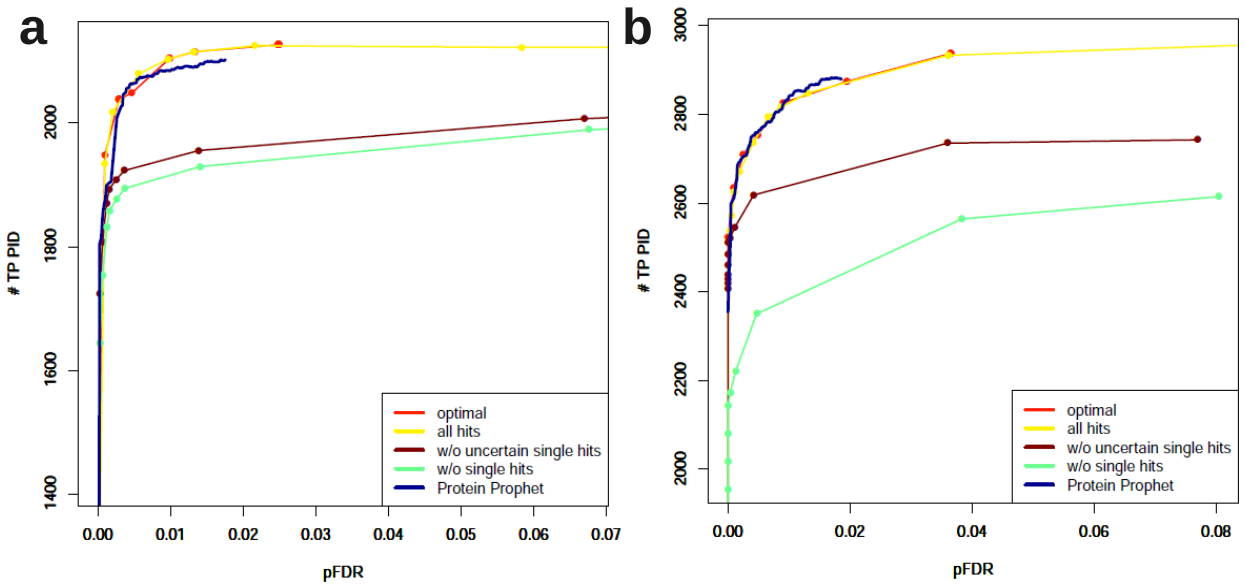

Figure 5: Optimal pruning strategies for various datasets. Expected number of *true positive* protein identifications (# TP PID) at varying protein identification false discovery rate (pFDR) levels are shown. Multidimensional fractionation high mass accuracy dataset for *L. interrogans* **(a)** and *S. pombe* **(b)** sample. Selection schemes considering all peptide-spectrum matches of sufficiently low peptide-spectrum match false discovery rate (yellow) effectively achieve the performance of the optimal selection scheme (red). Two alternative selection schemes are shown exemplarily. Selection schemes consequently neglecting single hits (green) or solely neglecting single hits with nonzero uncertainty (brown) achieve significantly suboptimal performance. The results for ProteinProphet (blue) are competitive with those of achieved with the optimal selection scheme.